Dermatol Ther (Heidelb)
https://doi.org/10.1007/s13555-024-01283-0

ORIGINAL RESEARCH

Automatic Acne Severity Grading with a Small and Imbalanced Data Set of Low-Resolution Images

Rémi Bernhard · Arnaud Bletterer · Maëlle Le Caro · Estrella García Álvarez · Belchin Kostov · Diego Herrera Egea

Received: August 9, 2024 / Accepted: September 19, 2024

ABSTRACT

***Introduction*:** Developing automatic acne vulgaris grading systems based on machine learning is an expensive endeavor in terms of data acquisition. A machine learning practitioner will need to gather high-resolution pictures from a considerable number of different patients, with a well-balanced distribution between acne severity grades and potentially very tedious labeling. We developed a deep learning model to grade acne severity with respect to the Investigator's Global Assessment (IGA) scale that can be trained on low-resolution images, with pictures from a small number of different patients, a strongly imbalanced severity grade distribution and minimal labeling.

***Methods*:** A total of 1374 triplets of images (frontal and lateral views) from 391 different patients suffering from acne labeled with the IGA severity grade by an expert dermatologist were used to train and validate a deep learning model that predicts the IGA severity grade.

***Results*:** On the test set we obtained 66.67% accuracy with an equivalent performance for all grades despite the highly imbalanced severity grade distribution of our database. Importantly, we obtained performance on par with more tedious methods in terms of data acquisition which have the same simple labeling as ours but require either a more balanced severity grade distribution or large numbers of high-resolution images.

***Conclusions*:** Our deep learning model demonstrated promising accuracy despite the limited data set on which it was trained, indicating its potential for further development both as an assistance tool for medical practitioners and as a way to provide patients with an immediately available and standardized acne grading tool.

***Trial Registration*:** chinadrugtrials.org.cn identifier CTR20211314.

R. Bernhard (✉) · A. Bletterer · M. Le Caro
QuantifiCare, 06410 Biot, France
e-mail: rbernhard@quantificare.com

A. Bletterer
e-mail: abletterer@quantificare.com

M. Le Caro
e-mail: mlecaro@quantificare.com

E. García Álvarez · B. Kostov · D. Herrera Egea
Almirall 08980 Sant Feliu de Llobregat, Barcelona, Spain
e-mail: estrella.garcia@almirall.com

B. Kostov
e-mail: belchin.kostov@almirall.com

D. Herrera Egea
e-mail: diego.herrera@almirall.com

Published online: 08 October 2024

Keywords: Automatic acne severity grading; Machine learning; Standardized method; Data scarcity; Low-resolution

Key Summary Points

We develop a machine learning model that performs automatic acne severity grading.

Our method to train this model alleviates numerous data acquisition burdens: it can be trained with low-resolution images, a small data set, a highly imbalanced acne severity grade distribution in the data set, and a basic labeling (no localization of acne lesions is needed).

Our method could provide medical practitioners with a tool that is easy to implement as it alleviates constraints related to the data acquisition process, which is crucial for the development of telemedicine.

INTRODUCTION

Acne vulgaris is one of the most widespread skin conditions arising from an inflammatory disorder of the pilosebaceous unit, therefore appearing in high density areas of pilosebaceous units, namely the face, neck, upper chest, shoulders, and back. Four key factors are responsible for this condition: increased sebum production, hyperkeratinization of the follicular infundibulum, inflammation, and *Cutibacterium acnes*. Features of acne vulgaris consist of excess grease, non-inflammatory lesions (open and closed comedones), and inflammatory lesions (papules, pustules, nodules, and cysts) [1].

A vast proportion of young adults and teenagers are suffering from acne. For example, up to approximately 85% of people from 12 to 25 years old in the USA have acne [2]. Moreover, acne is a widespread condition among older adults as well with 15.3% of women and 7.3% of adults who are 50 years old and over reporting suffering from acne [3]. Overall, studies indicate that 95% of people are affected by acne vulgaris at some point in their lives [4].

Since acne often appears on visible parts of the body, it may affect a person's mental health. In fact, people suffering from acne report higher rates of anxiety [5] as well as other mental issues such as suicidal ideation in adolescents [6].

The significant economic impact of acne cannot be ignored with an estimated cost of more than three billion dollars per year in the USA for treatment and loss of productivity [7].

In view of both the psychological and economic consequences of acne vulgaris, it is essential for people to be able to get this skin condition diagnosed quickly and early. Notably, knowledge of the severity of acne is necessary to decide which treatment a patient must undergo. However, diagnosis of the severity of acne requires a dermatologist to be available and for the patient to be able to travel to the dermatologist's office. These two conditions delay the time for acne sufferers to be diagnosed and prescribed timely treatment, thereby potentially impacting their mental health and resulting in economic damage for society.

A solution to overcome dermatologist availability and location is to take advantage of automatic diagnosis methods. A person suffering from acne would simply need to take pictures of the affected area, send these pictures for an automatic assessment, and receive almost instantly a response. This process does not depend on the availability or the location of a dermatologist. Moreover, automatic diagnosis methods are fully standardized and reproducible. Indeed, another major drawback of on-site visits with dermatologists is the observed low agreement between dermatologists related to acne severity grading, even in a very simple setting, meaning that the same patient seen by different dermatologists may receive different opinions. In a study conducted by Beylot et al. [8], 10 photographs corresponding to clear representations of three grades of acne (mild, moderate, and severe) were presented to eight expert dermatologists. Of the 10 photographs, only two received the same grade from the eight dermatologists, seven photographs received two adjacent scores, and one photograph received the three possible grade scores.

Developing automatic acne severity grading methods is therefore of primary interest. To do so, one may want to leverage deep learning algorithms to benefit from state-of-the-art performance in the vision domain. However, developing a deep learning method for severity grading can be a costly endeavor.

Firstly, one may want a high-quality data set on which to train and evaluate the model. For acne severity grading, this means gathering high-resolution images from multiple patients with different acne severity with all the associated burdens. The objective of training a model on high-resolution images is that it will hopefully perform better as it should take advantage of more details.

Secondly, one must perform labeling, with one or more experienced dermatologists, for all the images with a specific severity grade or with tedious labeling such as identifying different types of acne lesions for each image.

Lastly, one must collect data in a way such that images cover all the range of severity grades ensuring a relatively balanced distribution, which can be challenging. A poorly balanced distribution may be the cause of poor adaptation capacities of the model on unseen images. The unbalanced distribution of the training data can bias the model towards some specific grades, resulting in the model overpredicting these grades on new unseen images.

Many scales have been developed for grade acne severity such as the IGA scale (Investigator's Global Assessment) [9] or GAGS scale (Global Acne Grading System) [10] which considers lesion counting. For the GAGS scale different face regions and different lesions are weighted distinctively and results in four different severity grades. The Hayashi scale [11] is based on counting acne inflammatory lesions and also results in four different grades. The GEA scale (Global Evaluation Acne) [12] is based on an overall evaluation of acne severity, similar to the IGA scale, and results in six different acne severity grades. Other specific grades such as the one following Chinese guidelines for acne vulgaris [13] also rely on an overall appreciation of acne severity grade and include four different severity grades.

Some work has already been conducted on the prediction of acne severity by using artificial intelligence. Wang et al. [14] considered a combination of the GAGS and Hayashi scale. Images were labeled with bounding boxes around each acne lesion. The authors used a regression model to predict the number of specific lesions (comedones, papules, pustules, nodules) from which a severity grade is derived. Alzahrani et al. [15] considered the Hayashi criterion. The authors leveraged a model (based on the Unet architecture [16]) to predict the acne lesion count that is used to determine the acne severity grade. Seité et al. [17] considered the GEA scale. The authors used a classification model to predict the acne severity grade for each triplet of images. Wu et al. [18] consider the Hayashi criterion. The authors trained a classification model that takes advantage of both ground-truth severity grade and total lesion counts. Yang et al. [19] considered a grading scale based on the Chinese guidelines for the management of acne vulgaris. The authors fine-tuned an InceptionV2 [20] model pre-trained on ImageNet [21] to predict the acne severity grade. Li et al. [22] considered a grading system that takes into account both primary lesions (comedones, papules, nodules) and signs of change (postinflammatory hyperpigmentation, and scarring). Records were taken with the VISIA[1] system. The authors trained a ResNet50 [23] model that predicts acne severity scores by considering both severity cores and acne lesion-bounding boxes.

Some authors have considered the IGA scale while developing automatic methods to perform acne severity grading. These methods are the ones it makes sense to compare against.

Lim et al. [24] considered a data set composed of 472 facial photography images from 416 patients. Each image was labeled by two student trainees and one researcher, and their grading was then combined by majority vote, with the supervision and verification of a dermatologist. The authors trained different deep learning models to predict the acne severity grade directly from the image. The overall method takes input

[1] https://www.canfieldsci.com/imaging-systems/visia-complexion-analysis/.

images of size 600 × 800 pixels to 1200 × 1600 pixels depending on the model.

Zhao et al. [25] trained a model with 1000 selfie images and evaluated it on 230 other images. All images were labeled with the help of 11 dermatologists. The authors extracted skin patches from an image corresponding to the forehead, both cheeks, and the chin. Preprocessing was then performed on these skin patches before giving them as input to a four-layer regression model on top of a Resnet50-based feature extractor to predict the acne severity grade. The overall method takes as input images of size 228 × 228 pixels.

Huynh et al. [26] considered the IGA scale. The data set consisted of 1572 images taken with a mobile phone (Apple or Android) at approximately 20 cm from the person's face. Each image was labeled with bounding boxes (a bounding box around each acne lesion: whitehead, blackhead, papule, pustule, cyst, and acne scar) and the severity grade. Labeling was performed first by a junior dermatologist, whose labeling was then reviewed by a senior dermatologist. The authors trained a detection model (Faster R-CNN [27] with Resnet50 backbone) to detect acne lesions in an image. For each image, the resulting acne lesion count was then given as input to another machine learning model that predicts the acne severity grade. The overall method takes as input images of size 224 × 224 pixels.

Our objective in this work is to present a method to automatically grade acne severity that can be developed with few low-resolution images that do not present a balanced distribution with regards to acne severity grades. We propose a deep learning-based automatic severity grading method which alleviates all the constraints related to data acquisition mentioned earlier; specifically, the need to use a large data set of high-resolution images with the most balanced grade distribution as possible and tedious labeling.

Specifically, we present a deep learning model that can be trained on:

- Low-resolution images
- A small number of images
- A significantly imbalanced distribution of severity grades
- Severity grade labeling only (no acne lesions localization)

METHODS

Acne Severity Scale

In this work, the considered acne severity scale is derived from the Investigator's Global Assessment of severity, recommended and approved by the US Food and Drug Administration (FDA) since 2005 [9] and then used in clinical trials and controlled experimental studies. The IGA relies on a global evaluation by a dermatologist based on the presence of dominant lesions and the extent of inflammation [28]. This scale considers five possible grades for acne severity: 0, clear; 1, almost clear; 2, mild, 3, moderate; and 4, severe. The different criteria used to assign these grades are detailed in Table 1.

As we can in see in Table 1, the IGA scale is subject to interpretation as there are no quantitative measurements of the number of noninflammatory and inflammatory lesions for each scale.

Data Collection

The data used in this study included pictures of patients acquired within the scope of an acne study. The goal of this study was to examine the effect of 1.5 mg/kg per day of sarecycline on acne vulgaris. The study involved 391 Chinese patients; 262 of them were given sarecycline and 129 were given placebo. The majority (87%) of patients were over 18 years old and 13% were between 9 and 17 years old. Overall, the mean age was 21.9 years old, with a standard deviation of 5 years. The youngest patient was 9 years old and the oldest was 45 years old. Patients weighed between 33 and 136 kg; 58.6% of patients were women, 41.4% are men.

The study protocol required each patient to come to a clinical center for five successive visits, all separated by 3 weeks. As a result of logistical constraints some patients could not attend the

Table 1 IGA scale criteria

Score	Grade	Description
0	Clear	No evidence of papules or pustules
1	Almost clear	Rare: inflammatory papules (papules must be resolving and may be hyperpigmented, though not pink-red)
2	Mild	Few: inflammatory lesions (papules/pustules only; no nodulocystic lesions)
3	Moderate	Multiple: inflammatory lesions present; many papules/pustules; there may or may not be a few nodulocystic lesions
4	Severe	Inflammatory lesions are more apparent, many papules/pustules; there may or may not be a few nodulocystic lesions

five visits, producing a varying amount of visits for each patient.

For each visit, an experienced dermatologist rated the acne severity with respect to the IGA scale by looking at the patient directly. The dermatologist could palpate lesions to better identify them. Subsequently, three pictures were taken: one front face, one left profile, and one right profile. Participants with acne severity beyond grade 4 were not enrolled for the study.

Pictures were taken with iPhone 11 with a ring light and a resolution of 3024×4032 pixels.

This study was performed in accordance with the Helsinki Declaration of 1964 and its later amendments. The study was reviewed on March 23, 2021 by the Ethics Commitee of Peking University People's Hospital, No. 11, Xizhimen South Stree, Xicheng District Beijing, acting as a central ethics committee for all participating centers. Patients, parents, or guardians agreed to three pictures of the face (frontal, right side, and left side) being taken per visit for documenting acne distribution and by signing an informed consent.

We consolidated IGA grades 0 and 1 and grades 3 and 4 together, thus resulting in three groups of grades: 0–1, 2, and 3–4. This grouping arises from clinical considerations: grouped grades correspond to similar treatment. This grouping is commonly used in the literature of acne severity grading for this reason [24].

Therefore, the total data set consisted of 1347 triplets of images each of which was labeled according to the IGA acne severity grade.

Figure 1 shows three example images of a person presenting a certain number of noninflammatory lesions on the cheeks and a certain number of inflammatory lesions on the forehead and cheeks. After an onsite examination of the patient, an experienced dermatologist assigned the IGA severity score of 3 (moderate).

Our data set has two notable points on which we will elaborate further.

Firstly, there is a strongly imbalanced distribution of acne severity grades for our data set for all visits (Fig. 2): there was a majority of grades 3–4 and very few images of grade 0–1 (12.32% of the total number of visits). When training a deep learning model to predict the acne severity grade based on the triplets of images, this imbalance will make generalization to grade 0–1 (and to a lesser extent to grade 2) more difficult than in a balanced set as the model will tend to focus on images with grade 3–4.

Secondly, there are 391 patients and a total of 1347 visits. This data set size is considered to be small compared with similar machine learning problems. Relative to the image acquisition protocol of the study, the same person may thus be represented in up to five triplets of images. In our data we therefore have some triplets of images corresponding to different acne severity grades (up to three) but the same identity, corresponding to the evolution of the disease on the same patient between different visits. When training a deep learning model to predict the severity grade based on a triplet of images, these triplets with different grades but the same identity will make the task more difficult and

Fig. 1 Pictures taken during a visit of a patient assigned IGA severity grade 3 (moderate)

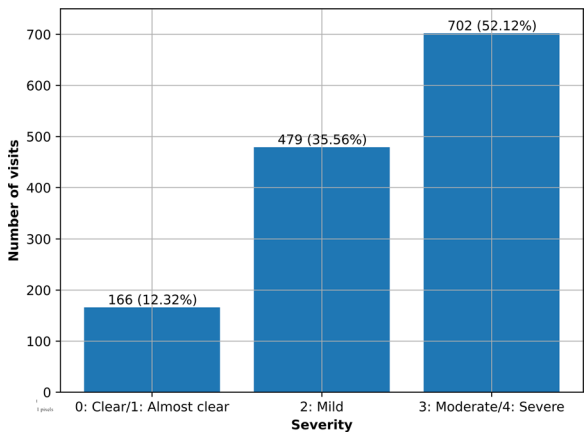


Fig. 2 Distribution of severity grades for our data (all visits)

induce confusion for the model. In fact, some acne lesions may be present on all the triplets but at different stages. For example, an acne lesion may be present on two images but for one of them it is considered as healed by the dermatologist, even if it leaves a red zone on the patient's face. During the training of the deep learning model, this information confuses the model as it appears as similar visual information at the same location but with different grades. An illustration of this phenomenon is presented in Fig. 3.

These important aspects of the data set, imbalanced grade distribution and a low number of patients, challenge the development of a robust deep learning model, given such methods typically require larger data sets to achieve reliable and generalizable performance to reduce potential bias in the learning process.

Model Architecture

We considered the ResNet architecture for our deep learning model, more precisely the ResNet110 with 1.7 M tunable parameters. As we aimed to developed a model that can be trained only on low-resolution images to alleviate burdens related to data acquisition, we chose a classic input shape for this model of size 224×224 pixels.

Pre-processing

Images that are fed into our deep learning model must be of size 224×224 and therefore images must be downsized to that dimension to be processed by the model. Before resizing images, we performed pre-processing to keep the most useful information possible. As seen in Fig. 1, a large part of the image consists of background which is considered useless for our task. We therefore

Fig. 3 Three face pictures corresponding from left to right to grades 3–4, 2, and 0–1. Some lesions are healed but still appear red (e.g., the area of the cheek just below the eye), causing confusion when training a deep learning model

performed face detection on each image to detect and crop out of the image only the region corresponding to the face of the patient. Face detection was performed via the BlazeFace Mediapipe solution[2] that offers a model tailored for detection of faces within 2 m from the camera. An example of an original image with its cropped version is presented in Fig. 4, wherein we can see that only relevant information is kept. After the face of the patient is cropped, resizing to size 224×224 is performed.

Training Procedure

We split the 1347 visits into a training, validation, and testing set of 1077, 135, and 135 triplets of images, respectively, corresponding to the usual split adopted by machine learning practitioners, i.e., 80%, 10%, and 10% of the total number of visits for the training, validation, and testing set, respectively. The training set is used to train the machine learning model, the validation set to fine-tune model parameters, and the testing set to eventually assess its real performance on unseen data. The final model is obtained after training and fine-tuning on the training set and validation set, respectively. This model performance is then evaluated on the testing set.

We were careful, among each set, to keep the closest distribution of severity grades to that of the initial data. More precisely, we randomly split the 1347 visits for the best evaluation possible while being attentive to the number of visits of each grade in each set. This is done by successive random samplings of the entire data set while checking the distribution of severity grades in each set. The resulting repartition is presented in Table 2.

When training our deep learning model to predict acne severity grades, considering the particularities of our data set, we must increase our model generalization capacity with additional method specifically for grade 0–1. First, we are in a setting where only a few images are available to train our model, which will tend to overfit to these images, and thus cause overfitting if nothing specific is done. Second, the strong imbalance of the acne severity grade distribution makes generalization even harder for grade

[2] https://github.com/google/mediapipe/blob/master/docs/solutions/face_detection.md.

Fig. 4 Original image and its cropped version

Table 2 Distribution of acne severity grades within each set

Data set	Grade 0–1	Grade 2	Grade 3–4
Full (1347)	12.32% (166)	35.56% (479)	52.12% (702)
Training (1077)	12.72% (137)	35.38% (381)	51.90% (559)
Validation (135)	11.11% (15)	40.74% (55)	48.15% (65)
Testing (135)	10.37% (14)	31.85% (43)	57.78% (78)

0–1 for which a small minority of images are available: during training, the model will tend to focus on giving a good prediction almost only for grade 3–4 as images for this grade represent almost all the data.

To manage the two aforementioned issues, we were required to deploy different strategies. We used data augmentation during training to increase the generalization capacity of our model. During training, each time a pair consisting of a triplet of images and an acne severity grade was presented to our model, each image of this triplet was modified randomly with various transformations. By doing so, it increased the diversity of images that our model observed during training, promoting its generalization capacity. Specifically, we considered four possible transformations: vertical flip, horizontal flip, rotation, and brightness change. Each image could either remain unmodified with a 20% probability, or undergo one of the four transformations described with a 20% probability each.

Then, we used dropout [29] to improve generalization, which is a regularization method that consists in setting parameters of a model to 0 randomly during training to prevent overfitting.

Finally, to promote generalization for grade 0–1 (and to a lesser extent for grade 2), we used a weighted version of the cross-entropy loss. Cross-entropy loss is a canonical loss function used to train deep learning models when considering classification tasks. During training, we weighted this loss function to penalize the model more when it does not predict well a triplet of images with ground-truth label 0–1, to make the model focus more on grade 0–1. The model was also penalized but to a lesser extent when it does not predict well a triplet of images with ground-truth label 2.

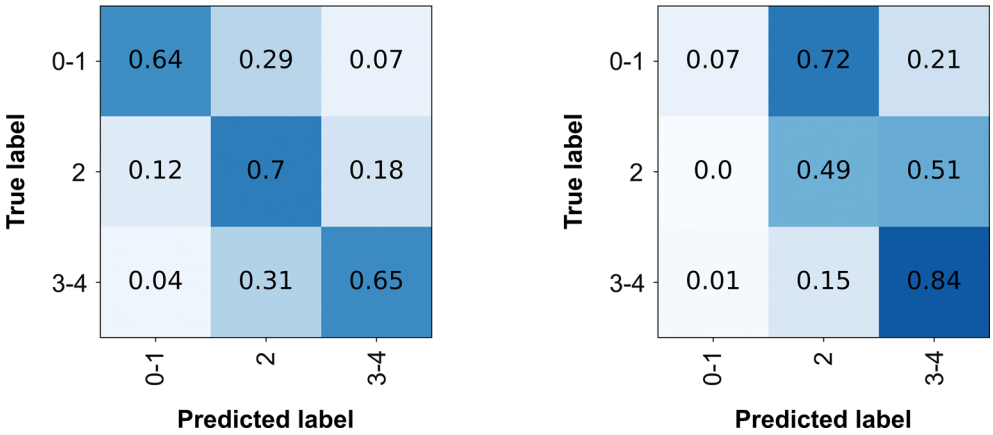


Fig. 5 Confusion matrix for our deep learning model with weighted loss (left) and for the same model trained without weighted loss (right). Percentage equals 100% for each row (each grade)

We trained our model for 1000 epochs, a batch size of 12, and the Adam optimizer starting with learning rate 0.01, decreasing to 0.001 after 800 epochs. All experiments were performed with a GPU Nvidia GeForce RTX 3080.

The number of epochs indicates how many times the training set is shown to the model to train it. The batch size refers to how many examples of the data set are shown to the model at a time during an epoch.

The model was trained on the training set, early stopping is performed relative to the performance on the validation set, and then evaluated on the testing set.

Metrics

We evaluate our model with respect to different metrics and criteria. We will look at the global accuracy of our model as well as the recall rate, precision rate, and F1 score for each of the three grades. It is crucial to look at each grade independently to ensure that there is no discrepancy for performance between grades. Notably, we will pay attention to metrics regarding images for grade 0–1 (and grade 2 to a lesser extent) as it is an underrepresented grade, and we want our model to generalize well for this grade as well. It must be noted that recall for each grade corresponds to the accuracy considering only images of this grade.

In terms of interpretability, the higher the recall metric for a grade, the more the model can predict accurately this grade when shown this grade. The higher the precision metric for a grade, the higher the model can predict accurately this grade among all grades.

RESULTS

The results below were obtained with our model and are used to illustrate the pertinence of our weighted loss to make the model more generalizable for all labels, despite the imbalanced grade distribution. We present results with and without the weighted loss.

Figure 5 shows the confusion matrix on our testing set for our deep learning model and for the same model but trained without weighted loss.

Tables 3 and 4 demonstrate the precision, recall, and F1 score results for our deep learning model and for the same model but trained without weighted loss, respectively.

The global accuracy equals 66.67% and 64% for our model and a model trained without weighted loss, respectively.

We note in Fig. 5 that the model trained without weighted loss focuses almost only on grade 3–4 for which there is the most data, while

Table 3 Precision, recall, and F1 score for the three grades for our deep learning model trained with weighted loss

	Grade 0–1	Grade 2	Grade 3–4
Precision	0.53	0.52	0.85
Recall	0.64	0.70	0.65
F1 score	0.58	0.59	0.74

our model is much more balanced between grades in terms of results. This illustrates the fact that our model allows for better generalization despite the imbalanced grade distribution.

Moreover, we see in Tables 3 and 4 that our model gives an accuracy per class (represented by the recall metric) that is similar for all three grades despite the highly unequal number of samples for each grade. The difference between grade 3–4 and the two other labels in terms of F1 score and precision is not as pronounced as the difference in terms of distribution between grade 3–4 and the two other labels. These observations do not hold when considering the model trained without weighted loss. Indeed, for this model, as we see in Table 4, the accuracy per class is highly unequal, with almost 0% accuracy for grade 0–1 and 83% accuracy for grade 3–4. For this model the F1 score also illustrates the trend of this model to focus predominantly on grade 3–4.

All these observations validate the relevance of our weighted loss to have a model that generalizes well to all labels, especially labels for which a small number of images are available because of an imbalanced grade distribution.

Figure 6 presents examples of triplets of images for which our model does not predict the correct grade. The boundary between grades may be fuzzy.

Table 4 Precision, recall, and F1 score for the three grades (without weighted loss)

	Grade 0–1	Grade 2	Grade 3–4
Precision	0.5	0.49	0.72
Recall	0.07	0.49	0.83
F1 score	0.13	0.48	0.77

DISCUSSION

We have developed a deep learning model to automatically predict IGA grades from images. Importantly, we were able to obtain this model based on low-resolution images, with only a few images with few different patients and with a strong imbalance in the grade distribution. Eventually, our model can accurately predict the severity grade for low-resolution input images even for grades where only a minority of images were available during training.

We now compare this model to other works that considered the IGA scale.

Some work cannot be directly compared to ours as they are not in the same setting of data scarcity or low-cost labeling. For instance, the model developed by Huynh et al. [26] required four dermatologists who labeled both the severity grade and the bounding boxes around acne lesions. Moreover, 1452 different patients were available for this study.

Comparison with Models with Higher-Resolution Input Images

Lim et al. [24] considered the same IGA grades in their work and are in a setting of data scarcity with only 416 different patients (vs 391 for our work). The repartition of their training data for grades is presented in Table 5.

Contrary to our data where we have most images with a grade 3–4 (52.12%), the authors considered data with an equal distribution between grades 0–1 and 2 and only 16.24% for grade 3–4. Their training data distribution is slightly more uniform than ours.

Figure 7 presents the confusion matrix on our test set (135 images of size 224 × 224 pixels) and Lim et al.'s best confusion matrix on their test set (98 images of size 1200 × 600 pixels) [24]. The global accuracy reached on their test set with their model equals 67% (equivalent to as ours).

In Fig. 7 we note that we have equivalent results to those of Lim et al. on our respective test sets. In fact, we have comparable results for grade 0–1 (first line of the matrix), slightly better results for grade 2, while Lim et al. have slightly better results for grade 3–4 (third line). However,

Fig. 6 Triplets of images for which the grade is not correctly predicted (top) ground truth 0–1, predicted 2; (middle) ground truth 2, predicted 0–1; (bottom) ground truth 2, predicted 3–4. The boundary between grades may be fuzzy

Table 5 Distribution of acne severity grades for the training set [24]

Data set	Grade 0–1	Grade 2	Grade 3–4
Full (472)	41.72% (131)	42.04% (132)	16.24% (51)

we focus on this comparison to highlight these results relative to the resolution they use to train their model and input images fed to the model, which equals 1200 × 600 pixels, and is thus far higher than ours (224 × 224 pixels). This higher input resolution plays an important role when considering the task of acne severity grading as it affects the precision of the information the model can rely on. Lim et al. showed that when considering a ResNet model and an input resolution of 600 × 800 pixels, the performance decreases. We show their confusion matrix on their test set for this model and resolution in Fig. 8. The global accuracy reached on their test set with this model equals 64%. For this resolution that stays far superior to ours, we observe that we have comparable performance on our test set for grades 0–1 and 3–4 and slightly better results for grade 2.

This comparison with the work of Lim et al. clearly illustrates the pertinence of our approach regarding the resolution of images considered. Indeed, our equivalent performance with a much lower input resolution

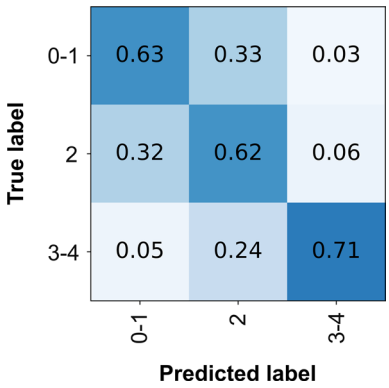
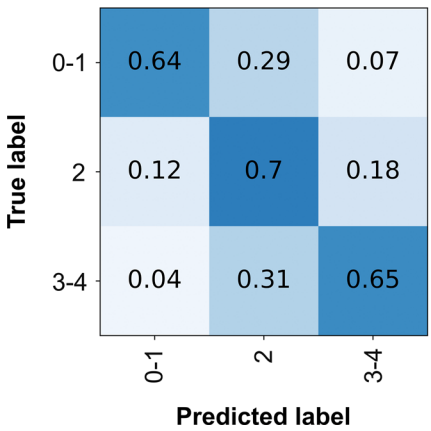


Fig. 8 Confusion matrix from Lim et al. [24] on their test set (input resolution equals 600 × 800 pixels). Percentage equals 100% for each grade (each row)

demonstrates one benefit of our method: it allows one to deploy a model that can take as input low-resolution images with similar performance to other approaches with high-resolution images. Importantly, this means that our model can be trained with low-resolution images, thereby alleviating numerous constraints associated with data acquisition.

Comparison with Models with a More Balanced Grade Distribution

Zhao et al. [25] considered five grades that correspond exactly to the IGA grades and are in a



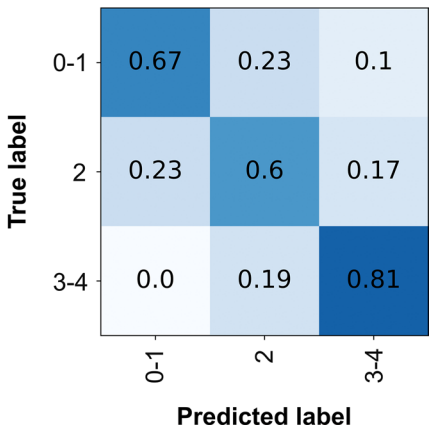


Fig. 7 (Left) Our confusion matrix on our test set (input resolution equals 224 × 224 pixels). (Right) Best confusion matrix from Lim et al. [24] on their test set (input resolution equals 1200 × 600 pixels). Percentage equals 100% for each grade (each row)

Table 6 Distribution of acne severity grades for the training set [25]

Data set	Grade 0–1	Grade 2	Grade 3–4
Training (1000)	37% (370)	43.00% (430)	20.00% (200)

Table 7 Precision, recall, and F1 score for the three grades [25]

	Grade 0–1	Grade 2	Grade 3–4
Precision	0.37	0.62	0.96
Recall	0.33	0.82	0.39
F1 score	0.35	0.71	0.55

setting similar to ours. They had a data set consisting of 1230 images labeled with the IGA scale and an input resolution for their deep learning model of 224 × 224 pixels. Table 6 presents the estimated distribution of acne severity grades for their training set (based on Fig. 1 in their article). We can clearly see that the distribution of acne severity grades for their data is more balanced than ours (Table 2), as it is closer to the uniform distribution. Notably, their training set has an almost equal number of images for grades 0–1 and 2.

In their work, the authors considered the five original IGA grades. To compare their work fairly with ours, we converted their results to only represent the three IGA grades that we consider. Figure 9 presents the confusion matrix on our test set (135 images of size 224 × 224 pixels) and their confusion matrix on their test set (230 images of size 224 × 224 pixels).

Table 7 presents the precision, recall, and F1 score for their model.

Compared to our model that shows almost similar performance for all grades (Table 3) on our test set, we see that Zhao et al.'s model focuses much more on grade 2 for their test set. The accuracy per label (recall) equals 82% for grade 2 but only 33% and 39% for grades 0–1 and 3–4, respectively. Moreover, their confusion matrix shows that their model is more likely to predict grade 2 than any other grade. Our model is able to generalize to all severity grades despite our grade distribution being less balanced than theirs. The comparison with the work of Zhao et al. clearly illustrates the relevance of our method to obtain a model that generalizes equally well among acne severity grades, almost independently from the grade distribution in data. This is particularly important as the process of acquiring data with a balanced grade distribution can be both tricky and tedious.

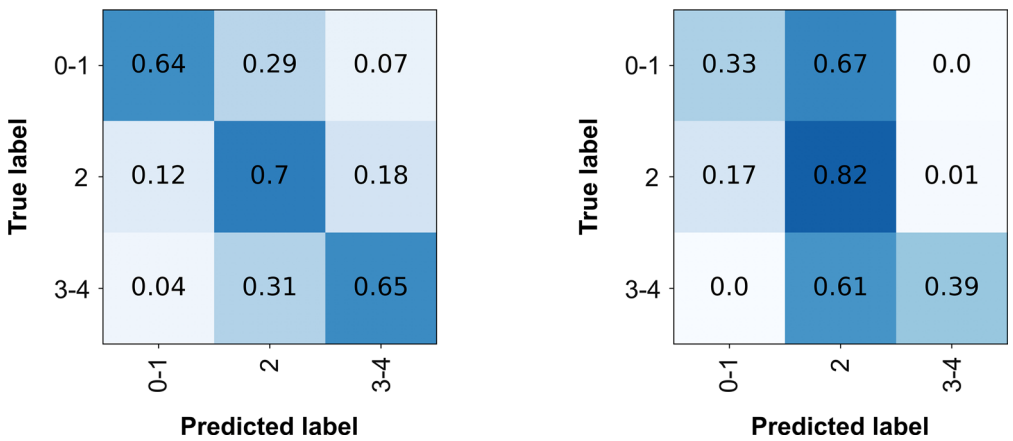


Fig. 9 (Left) Our confusion matrix on our test set. (Right) Confusion matrix from Zhao et al. [25] on their test set. Percentage equals 100% for each grade (each row)

Perspectives and Limitations

Our method can be improved in different ways, each of which fits our data scarcity and label distribution imbalance setting.

First, we could leverage a model pre-trained on a similar task to perform transfer-learning [30], which consists in using part of the pre-trained model weights as a starting point for the training on our data. Transfer learning has been shown to enhance the generalization capacity of the final model.

Second, we could explore self-supervised learning [31], a method that utilizes unlabeled data to train a backbone model, to which additional layers can be added and the entire model fine-tuned on our data. In some cases, self-supervised learning attains better performance than the supervised setting. For our purposes, self-supervised learning would be applied to unlabeled acne data.

Beyond these performance improvements, future work could involve developing a model to predict acne evolution of a patient.

We identified two limitations to our work that could be interesting to study in future work. The first limitation is that for all image-based methods, the model has to make decisions on the basis of images only and does not have access to information than can be obtained through other means. In our case, the model does not have access to information that can be obtained via on-site examination, like palpation for example. This is the main limitation of an image-based system. The second limitation has to do with the labeling we use for our study. Regarding machine learning, we need ground-truth labels to train our model. As the IGA scale is qualitative, it is subject to interpretation. This means that the notion of "true" ground-truth labels does not exist in our case. The most important factor for our ground-truth labels is then consistency: we need to have consistency in ground-truth labels on the whole data set. Having our data labeled by only one dermatologist ensures this consistency. This consistency would not have held if multiple dermatologists had labeled different subsets of the data set. However, one limitation of this type of labeling is that our machine learning model ends up being biased towards the dermatologist interpretation of the IGA score. Having a data set of ratings aggregated between different dermatologists would have led to a model less biased towards one particular rater. Having this new type of labeling would be an interesting perspective to our work.

CONCLUSION

We propose a method that allows one to train a machine learning model that predicts IGA acne severity grades while alleviating numerous burdens associated with data acquisition. Notably, we present how to train the model with low-resolution pictures from a small number of different patients, with only severity grade labeling and a highly imbalanced severity grade distribution in training data.

In addition, a model trained with our method affords results which are equivalent to those obtained by other methods that consider a more involved data acquisition process. Notably, we compare our method to others that require high-resolution images or a well-balanced distribution of acne severity grades in the training data. Importantly, we verify that our model performs equally well for all grades.

The proposed method is essential for medical practitioners seeking an assistance tool without the associated constraints and the need for a significant data acquisition setup. With our methodology, a medical practitioner does not need to collect images from numerous different patients with a balanced grade distribution nor require sophisticated labeling such as lesion localization. Instead, only the severity grade associated with an image is necessary.

Another important positive impact of our method is related to telemedicine. The method developed in this work allows for clinical centers to be able to develop their own automatic grading tool more easily, and with consumer-grade devices like smartphones. This is particularly crucial in the context of telemedicine. Indeed, patients will be able to have their acne severity graded in an automatic and standardized

manner, without having to rely on practitioner availability.

ACKNOWLEDGEMENTS

We thank the participants of the study.

Author Contributions. Methodology/Experiments/Writing—original draft preparation: Rémi Bernhard. Writing – review and editing: Arnaud Bletterer, Maëlle Le Caro, Estrella García Álvarez, Belchin Kostov, Diego Herrera Egea. Data collection: Estrella García Álvarez, Belchin Kostov, Diego Herrera Egea. Supervision: Arnaud Bletterer, Estrella García Álvarez, Belchin Kostov, Diego Herrera Egea. We thank Dr. Aida Fernández Rubio for her review of the work as well as for helpful discussions.

Funding. No funds, grants or other support was received. The Rapid Service Fee will be paid by QuantifiCare.

Data Availability. The data that support the findings of this study are not openly available due to reasons of sensitivity and privacy.

Declarations

Conflict of Interest. The authors declare that they have no conflicts of interest.

Ethical Approval. This study was performed in accordance with the Helsinki Declaration of 1964 and its later amendments. The study was reviewed on March 23, 2021 by the Ethics Commitee of Peking University People's Hospital, No. 11, Xizhimen South Stree, Xicheng District Beijing, acting as a central ethics committee for all participating centers. Patients, parents, or guardians agreed to three pictures of the face (frontal, right side, and left side) being taken per visit for documenting acne distribution and by signing an informed consent.

REFERENCES

1. Williams HC, Dellavalle RP, Garner S. Acne vulgaris [published correction appears in Lancet. 2012 Jan 28;379(9813):314]. Lancet. 2012;379(9813):361–72. https://doi.org/10.1016/S0140-6736(11)60321-8.

2. Zaenglein AL. Acne vulgaris. N Engl J Med. 2018;379(14):1343–52. https://doi.org/10.1056/NEJMcp1702493.

3. Collier CN, Harper JC, Cafardi JA, et al. The prevalence of acne in adults 20 years and older [published correction appears in J Am Acad Dermatol. 2008 May;58(5):874. Cafardi, Jennifer A [added]]. J Am Acad Dermatol. 2008;58(1):56–9. https://doi.org/10.1016/j.jaad.2007.06.045.

4. Madden WS, Landells ID, Poulin Y, et al. Treatment of acne vulgaris and prevention of acne scarring: Canadian consensus guidelines. J Cutan Med Surg. 2000;4(Suppl 1):S2–13.

5. Ramrakha S, Fergusson DM, Horwood LJ, et al. Cumulative mental health consequences of acne: 23-year follow-up in a general population birth cohort study. Br J Dermatol. 2016;175(5):1079–81. https://doi.org/10.1111/bjd.13786.

6. Halvorsen JA, Stern RS, Dalgard F, Thoresen M, Bjertness E, Lien L. Suicidal ideation, mental health problems, and social impairment are increased in adolescents with acne: a population-based study. J Invest Dermatol. 2011;131(2):363–70. https://doi.org/10.1038/jid.2010.264.

7. Bhate K, Williams HC. Epidemiology of acne vulgaris. Br J Dermatol. 2013;168(3):474–85. https://doi.org/10.1111/bjd.12149.

8. Beylot C, Chivot M, Faure M, et al. Inter-observer agreement on acne severity based on facial photographs. J Eur Acad Dermatol Venereol. 2010;24(2):196–8. https://doi.org/10.1111/j.1468-3083.2009.03278.x.

9. FDA. Guidance for industry acne vulgaris: developing drugs for treatment. https://downloads.regulations.gov/FDA-1975-N-0012-0317/attachment_250.pdf. Accessed 10 Sep 2024.

10. Doshi A, Zaheer A, Stiller MJ. A comparison of current acne grading systems and proposal of a novel system. Int J Dermatol. 1997;36(6):416–8. https://doi.org/10.1046/j.1365-4362.1997.00099.x.

11. Hayashi N, Akamatsu H, Kawashima M, Acne Study Group. Establishment of grading criteria for acne severity. J Dermatol. 2008;35(5):255–60. https://doi.org/10.1111/j.1346-8138.2008.00462.x.

12. Dréno B, Poli F, Pawin H, et al. Development and evaluation of a Global Acne Severity Scale (GEA Scale) suitable for France and Europe. J Eur Acad Dermatol Venereol. 2011;25(1):43–8. https://doi.org/10.1111/j.1468-3083.2010.03685.x.

13. Chinese guidelines for the management of acne vulgaris: 2019 update. Int J Dermatol Venereol. 2019;2(3):129–138. https://doi.org/10.1097/JD9.0000000000000043.

14. Wang J, Luo Y, Wang Z, et al. A cell phone app for facial acne severity assessment. Appl Intell (Dordr). 2023;53(7):7614–33. https://doi.org/10.1007/s10489-022-03774-z.

15. Alzahrani S, Al-Bander B, Al-Nuaimy W. Attention mechanism guided deep regression model for acne severity grading. Computers. 2022;11(3):31. https://doi.org/10.3390/computers11030031.

16. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. https://arxiv.org/abs/1505.04597. Accessed 14 Dec 2023.

17. Seité S, Khammari A, Benzaquen M, Moyal D, Dréno B. Development and accuracy of an artificial intelligence algorithm for acne grading from smartphone photographs. Exp Dermatol. 2019;28(11):1252–7. https://doi.org/10.1111/exd.14022.

18. Wu X, Wen N, Liang J, et al. Joint acne image grading and counting via label distribution learning. In: IEEE/CVF International Conference on Computer Vision (ICCV). 2019; p. 10641–10650. https://doi.org/10.1109/ICCV.2019.01074.

19. Yang Y, Guo L, Wu Q, et al. Construction and evaluation of a deep learning model for assessing acne vulgaris using clinical images. Dermatol Ther (Heidelb). 2021;11(4):1239–48. https://doi.org/10.1007/s13555-021-00541-9.

20. Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015:2818–2826. https://doi.org/10.1109/CVPR.2016.308.

21. Deng J, Dong W, Socher R, et al. ImageNet: a large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015; p. 248–255. https://doi.org/10.1109/CVPR.2009.5206848.

22. Li J, Du D, Zhang J, et al. Development and validation of an artificial intelligence-powered acne grading system incorporating lesion identification. Front Med (Lausanne). 2023;10:1255704. https://doi.org/10.3389/fmed.2023.125570.

23. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016; p. 770–778. https://doi.org/10.1109/CVPR.2016.90.

24. Lim ZV, Akram F, Ngo CP, et al. Automated grading of acne vulgaris by deep learning with convolutional neural networks. Skin Res Technol. 2020;26(2):187–92. https://doi.org/10.1111/srt.12794.

25. Zhao T, Zhang H, Spoelstra J. A computer vision application for assessing facial acne severity from selfie images. https://arxiv.org/abs/1907.07901. Accessed 14 Dec 2023.

26. Huynh QT, Nguyen PH, Le HX, et al. Automatic acne object detection and acne severity grading using smartphone images and artificial intelligence. Diagnostics (Basel). 2022;12(8):1879. https://doi.org/10.3390/diagnostics12081879.

27. Ren S, He K, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell. 2017. https://doi.org/10.1109/TPAMI.2016.2577031.

28. Zarchi K, Jemec GBE. Severity assessment and outcome measures in acne vulgaris. Curr Derm Rep. 2012;1(3):131–6. https://doi.org/10.1007/s13671-012-0016-8.

29. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, et al. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res. 2021;15:1929–58.

30. Bozinovski S. Reminder of the first paper on transfer learning in neural networks. Informatica (Slovenia). 2021. https://doi.org/10.31449/inf.v44i3.2828.

31. Dosovitskiy A, Springenberg J, Riedmiller M, Brox T. Discriminative unsupervised feature learning with exemplar convolutional neural networks. IEEE Trans Pattern Anal Mach Intell. 2016;38:1734–47. https://doi.org/10.1109/TPAMI.2015.2496141.